Preventive Veterinary Medicine 116 (2014) 325–335

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Using open-access taxonomic and spatial information to create a comprehensive database for the study of Mammalian and avian livestock and pet infections☆

K.M. McIntyre[a,*], C. Setzkorn[a], M. Wardeh[a], P.J. Hepworth[a], A.D. Radford[b], M. Baylis[a]

[a] Department of Epidemiology and Population Health, Institute of Infection and Global Health (IGH), University of Liverpool (UoL), Leahurst Campus, Neston, Cheshire CH64 7TE, UK
[b] Department of Infection Biology, IGH, UoL, UK

ARTICLE INFO

Article history:
Received 5 March 2013
Received in revised form 21 June 2013
Accepted 3 July 2013

Keywords:
Database
Disease
Pathogen
Zoonosis
Emerging
Surveillance

ABSTRACT

What are all the species of pathogen that affect our livestock? As 6 out of every 10 human pathogens came from animals, with a good number from livestock and pets, it seems likely that the majority that emerge in the future, and which could threaten or devastate human health, will come from animals. Only 10 years ago, the first comprehensive pathogen list was compiled for humans; we still have no equivalent for animals. Here we describe the creation of a novel pathogen database, and present outputs from the database that demonstrate its value.

The ENHanCEd Infectious Diseases database (EID2) is open-access and evidence-based, and it describes the pathogens of humans and animals, their host and vector species, and also their global occurrence. The EID2 systematically collates information on pathogens into a single resource using evidence from the NCBI Taxonomy database, the NCBI Nucleotide database, the NCBI MeSH (Medical Subject Headings) library and PubMed. Information about pathogens is assigned using data-mining of meta-data and semi-automated literature searches.

Here we focus on 47 mammalian and avian hosts, including humans and animals commonly used in Europe as food or kept as pets. Currently, the EID2 evidence suggests that:

- Within these host species, 793 (30.5%) pathogens were bacteria species, 395 (15.2%) fungi, 705 (27.1%) helminths, 372 (14.3%) protozoa and 332 (12.8%) viruses.
- The odds of pathogens being emerging compared to not emerging differed by taxonomic division, and increased when pathogens had greater numbers of host species associated with them, and were zoonotic rather than non-zoonotic.
- The odds of pathogens being zoonotic compared to non-zoonotic differed by taxonomic division and also increased when associated with greater host numbers.
- The pathogens affecting the greatest number of hosts included: *Escherichia coli*, *Giardia intestinalis*, *Toxoplasma gondii*, *Anaplasma phagocytophilum*, *Cryptosporidium parvum*, Rabies virus, *Staphylococcus aureus*, *Neospora caninum* and *Echinococcus granulosus*.

* Corresponding author. Tel.: +44 151 794 6079; fax: +44 151 794 6005.
E-mail address: k.m.mcintyre@liv.ac.uk (K.M. McIntyre).

0167-5877/$ – see front matter
http://dx.doi.org/10.1016/j.prevetmed.2013.07.002

- The pathogens of humans and domestic animal hosts are characterised by 4223 interactions between pathogen and host species, with the greatest number found in: humans, sheep/goats, cattle, small mammals, pigs, dogs and equids.
- The number of pathogen species varied by European country. The odds of a pathogen being found in Europe compared to the rest of the world differed by taxonomic division, and increased if they were emerging compared to not emerging, or had a larger number of host species associated with them.

1. Introduction

A great deal of time and resources have been put into studying individual pathogens or groups of pathogens affecting the animals with which humans have the most contact, because they potentially affect food security including socio-economic impacts and zoonotic disease transmission. For instance, it is estimated that one in four people in the UK annually suffer from diarrhoeal disease (Tam et al., 2012) and as previous work has suggested that around 60% of infectious organisms known to be pathogenic to humans are zoonotic (Taylor et al., 2001; Woolhouse and Gowtage-Sequeria, 2005), it is likely that many of these episodes of illness are a result of preceding transmission from animals, causing emerging infections in humans.

Just over 10 years ago, the first attempt to produce an inventory of the pathogens of humans was undertaken (Taylor et al., 2001). At the same time, this list was combined with further information for livestock, dogs, cats and wildlife, in a study which aimed to quantify both pathogen characteristics and their interactions with key features of pathogen–host epidemiology including host range, zoonotic or emerging disease status and socio-economic importance (Cleaveland et al., 2001). The main material for both studies came from gathering specific information from textbooks; however, scientific literature was also examined to ascertain emerging infections. Such sources are potentially biased towards clinical infections and the data gathering would have been a lengthy process. There may be gaps in knowledge and therefore biases in animal and particularly companion animal disease data due to a lack of joined-up surveillance, which several recent projects aim to rectify (Moore et al., 2004a,b; Paiba et al., 2007; Radford et al., 2010). In addition, the job of identifying publications on pathogens and their hosts may be made more difficult due to our domestication of animals. For example, pathogens are found in domestic hosts in which they would not, naturally, be expected to occur, and husbandry practices can introduce host or pathogen vector species into new areas, change host susceptibility or behaviour, exposing hosts to new pathogens via modified transmission routes. Previous control programmes using, for instance antibiotics, can also promote further evolution of pathogens, changing their pathogenesis in host populations.

Within the present study, we describe the creation of an open-access pathogen database which was constructed as a part of the ENHanCE project (McIntyre et al., 2010), and present a number of outputs from the database to demonstrate its value as a tool. Within the results section, we focus on 47 mammalian and avian hosts, including humans and animals commonly used in Europe as food or kept as pets.

The open-access evidence-based ENHanCEd Infectious Diseases (EID2) database (University of Liverpool, 2011), provides a large, automated and systematically generated method of studying the main pathogens and hosts involved in disease transmission. It describes the pathogens of humans and animals, their host and vector species, and also their global occurrence, and the information contained within it is likely to reflect biases in the research undertaken on pathogens and their hosts.

The main way in which the results of our study differ from that of Cleaveland et al. (2001) is in the use of the EID2 source for pathogen information. All evidence within the EID2 comes from, and is linked to, previously published sources; the database extracts and analyses material contained in the meta-data of millions of nucleotide sequences and in publications, storing it in a hierarchical phylogenetic tree structure. As a result, where Cleaveland et al. (2001) used textbooks to find pathogens of specific hosts, in the EID2 approach the evidence comes from individual reports. The semi-automated nature of information gathering has also meant: a much larger quantity of proof of host–pathogen interactions has been used as evidence; this proof comes from primary, usually peer-reviewed literature; it has been possible to study many more pathogens if information has been published on them; and it has been possible to use a more exhaustive list of domestic animal hosts. In addition, spatial information for pathogens assigned at the country-level has been built into the EID2. Within this study we have examined differences in the evidence for the occurrence of pathogens in Europe compared to the rest of the world.

Our main aims were to: (1) Provide a description of the structure of the EID2 database including the data-sources used to create it. (2) Carry out an analysis demonstrating the usefulness of the EID2. This was achieved by comparing some of the content of the EID2 with the results of several earlier seminal papers. We show that the EID2 can be used to recreate these results potentially in a quicker, less biased, more easily repeatable and updateable way. The information it contains is a reflection of biases in the scientific research which is undertaken on pathogens and their hosts, however the database can be quickly updated when new information become available. Further, we emphasise that the EID2 is a much bigger resource which could be adapted to answer questions on other species, vector and pathogen assemblages, and on other drivers of disease.

2. Materials and methods

2.1. The ENHanCEd Infectious Diseases (EID2) database

2.1.1. Individual host, vector and pathogen information

Identification of pathogens was undertaken by creating an evidence-based database resource, The ENHanCEd Infectious Diseases database (EID2) (University of Liverpool, 2011), which stores evidence of the pathogens of all animals (not including fish), including humans, and their global occurrence. Its core is the NCBI Taxonomy database (National Center for Biotechnology Information, 2012c) which provides a hierarchical phylogenetic structure for host, vector or pathogen (here-after referred to as 'organism') nodes, such that outputs can be obtained for species and higher taxonomic groups (for instance "flaviviruses", "ruminants"). The information on each organism node includes alternative names (including synonyms), which were mostly provided as a part of the information from the NCBI Taxonomy database. If phylogenetic information for an organism was not included within the NCBI Taxonomy database, they were added manually into the EID2 at the correct place within the phylogenetic tree. Further information about each organism, such as their taxonomic rank (genus, species, etc.) or their taxonomic division for pathogens (bacteria – including rickettsia, fungi – including algal pathogens, helminths – including thorny-headed worms and pentastomids, protozoa, and viruses – including prion agents) is stored using a series of statements. These statements can be created using semi-automated methods such as data-mining of meta-data held within the NCBI Taxonomy database or the NCBI Nucleotide database (National Center for Biotechnology Information, 2012b), or they can be nominated by an individual using evidence from a publication. Data on publications described within PubMed (National Center for Biotechnology Information, 2012d) is also held in the EID2, with papers included based upon the organisms which they describe, and abstract information available for recently published papers. The EID2 has been set up in such a way that automated literature searches to look for specific topics associated with pathogens, for instance climate change, can be undertaken. In addition, the database is linked to climate data for a global grid ($0.25^\circ \times 0.25^\circ$), which can be used to model the spatial distribution of pathogens.

2.1.2. Pathogenic status of pathogens

In order to decide which pathogens cause significant clinical disease in hosts and therefore which need to be considered from a health and well-being perspective, a pathogenic status was assigned using expert opinion, to each pathogen node included within the EID2. This involved reviewing information from the literature on the clinical presentation of a disease caused by a pathogen within at least one of its host species. Definitions of pathogenic include:

- Frequently pathogenic – An organism which frequently has a clinically pathogenic effect (causes morbidity or mortality) upon humans or domestic animals.
- Non-pathogenic – An organism which causes no clinical signs within any of its hosts.
- Unknown pathogenicity – An organism for which there is insufficient evidence to decide whether it causes pathogenic effects in any host.

2.1.3. Information on pathogen–host and pathogen–location interactions

Specific information on pathogens affecting a certain host (termed a 'host–pathogen interaction') or pathogens occurring within a country (an 'organism–country interaction') was mined from meta-data held within the NCBI Nucleotide database (National Center for Biotechnology Information, 2012b); such information was treated as a 'gold standard' within the database. The data-mining was undertaken by searching the meta-data for entries describing infection of host species by pathogens (including bacteria, fungi, helminths, protozoa and viruses), or for entries describing pathogen infection occurring in hosts reported in a specific country, respectively. The last update from the nucleotide database was undertaken in December 2011. In addition, specific scientific publications were used as evidence of certain pathogen–host interactions.

A further source of information utilised for organism–country interactions came from automated searches of the PubMed database (National Center for Biotechnology Information, 2012d) and the NCBI MeSH (Medical Subject Headings) library (National Center for Biotechnology Information, 2012a); when the name of an organism and the (minor subject) MeSH term for a country co-occurred within a certain number of publications, an assumption was made about the occurrence of that organism within that country. Within the EID2, spatial data on organisms are hierarchically organised according to the NCBI MeSH library (National Center for Biotechnology Information, 2012a), thus allowing outputs at different regional-levels. Spatial outputs can be in the form of a list or a map.

2.2. Domestic animal species shortlist

Once the EID2 had been populated with information on organisms including pathogens, a short-list of humans and domestic animal hosts with which we have close contact in Europe was drawn up based on the agreement of experts involved in the ENHanCE project (McIntyre et al., 2010). Data for this host population was examined for the purposes of this study. This list included domestic animals we eat or companion animals we keep as pets, and exotic animals also used as food sources or as pets (Table 1).

2.3. Statistical analyses of data

2.3.1. The validity of using automatically mined pathogen–location information in the EID2 database

The threshold number of papers with which to infer a pathogen–location interaction from automated searches of the PubMed database (National Center for Biotechnology Information, 2012d) and the NCBI MeSH library (National Center for Biotechnology Information, 2012a) was investigated using two different approaches.

Table 1
Animal species including humans for which pathogens have been studied, including domestic animals we eat or companion animals we keep as pets, and exotic animals also used as food sources or as pets.

Scientific name	Common name	Scientific name	Common name
Agapornis personata	Masked lovebird	*Lama glama*	Lama
Agapornis roseicollis	Rosy-faced lovebird	*Lama pacos*	Alpaca
Anas platyrhynchos	Domestic duck	*Meleagris gallopavo*	Turkey
Anser anser	Domestic goose	*Melopsittacus undulatus*	Budgerigar
Bison bison	American bison	*Meriones unguiculatus*	Mongolian gerbil
Bison bonasus	European bison	*Mesocricetus auratus*	Syrian golden hamster
Bos indicus	Zebu	*Mus musculus*	House mouse
Bos taurus	Cow	*Mustela putorius furo*	Domestic ferret
Camelus dromedarius	Dromedary	*Numida meleagris*	Helmeted guineafowl
Canis lupus familiaris	Domestic dog	*Nymphicus hollandicus*	Cockatiel
Capra hircus	Domestic goat	*Oryctolagus cuniculus*	Domestic rabbit
Capreolus capreolus	Roe deer	*Ovis aries*	Sheep
Cavia porcellus	Domestic guinea pig	*Ovis aries musimon*	Mouflon
Cervus elaphus	Red deer	*Pavo cristatus*	Blue peafowl
Chinchilla lanigera	Chinchilla	*Phasianus colchicus*	Ring-necked pheasant
Columba livia	Domestic pigeon	*Rangifer tarandus*	Reindeer
Cricetus cricetus	Common hamster	*Rattus norvegicus*	Brown rat
Dama dama	Fallow deer	*Rattus rattus*	Black rat
Equus asinus	Domestic donkey	*Rhombomys opimus*	Great Gerbil
Equus caballus	Domestic horse	*Serinus canaria*	Canary
Felis catus	Domestic cat	*Struthio camelus*	Ostrich
Gallus gallus	Chicken	*Sus scrofa*	Wild boar
Homo sapiens	Humans	*Sus scrofa domesticus*	Domestic pig
Lagopus lagopus scotica	Red grouse		

First, the positive predictive value (PPV) of a putative pathogen–location interaction was assessed. Where automated searching suggested a positive interaction occurred, a randomly selected subset of papers (stratified according to the pathogen and continent on which the MeSH term country was located) was examined to see if there was supportive evidence for the interaction. We calculate the PPV as the proportion of predicted interactions for which papers provide supportive evidence (Thrusfield, 2007). We investigated whether the predicted interaction being supported by papers was affected by pathogenic status or taxonomic division using generalised linear models (GLM) with binomial errors and logit link functions or Chi-squared analysis, respectively. Within the sub-sample of papers used to examine a possible effect of taxonomic division, all papers provided supportive evidence for a pathogen–location interaction for the fungi, helminth and protozoa divisions, and so these were not included in the statistical analysis. If pathogen–location interactions had previously been described using our 'gold-standard' – information provided as a part of NCBI Nucleotide database meta-data (National Center for Biotechnology Information, 2012b), they were not included within this analysis.

Second, a GLM with binomial errors and a logit link function was used to ascertain the odds of an inferred pathogen–country interaction being correct. The outcome variable within the model was if a pathogen–country interaction had been reported in meta-data from the NCBI Nucleotide database for at least one nucleotide sequence. The explanatory variable was the number of papers from a PubMed search in which a pathogen and country MeSH term had co-occurred. Non-linear relationships in the data were investigated using generalised additive modelling (GAM) and inclusion of polynomial terms or a linear spline function. The break points within the spline function were explored using an iterative process where the break increased from the minimum in discrete steps representing the number of papers. Significantly improved GLM model fits were established by comparing models using Chi-squared analysis; the final GLM model was ascertained using deviance residuals with the smallest value being the best fit. The final GLM model included $\log_{10}(n+1)$ transformation of the covariate. Further to the linear spline model, a model in which the number of papers from PubMed was recoded as a factor was also tested, again exploring breaks using an iterative process. This technique would allow simple interpretation of the characteristics of each part of the relationship between nucleotide sequences and PubMed papers.

2.3.2. Pathogen range within hosts

GLMs with binomial errors and logit link functions were used to explore if the odds of a human pathogen being emerging was influenced by: the number of hosts it occurred in (one host, two hosts or more than two hosts) and whether the pathogen was zoonotic or non-zoonotic. Information on the taxonomic division of pathogens and their pathogenicity was also included within the analyses as covariates. Emerging and zoonotic statuses were taken from Taylor et al. (2001) and Woolhouse and Gowtage-Sequeria (2005); they were not available for every pathogen species. The models were built using stepwise deletion of terms by comparing models using Chi-squared analysis, and statistical significance was determined by a *P*-value of less than 0.05. Adjusted odds ratios (AOR) significantly different from one were used as an indicator of raised or lowered odds of pathogens being emerging. Hosmer–Lemeshow goodness-of-fit tests were used to judge the goodness-of-fit of the models. The influence of number of hosts and emerging status upon

the odds of pathogens being zoonotic compared to non-zoonotic was explored using models which included the same covariates and methodologies as described above for emerging pathogens.

2.3.3. Differences in the spatial representation of pathogen species

The influence of number of hosts, emerging and zoonotic status upon the odds of pathogens being found in Europe compared to the rest of the world was explored using models which used the same methodologies as described above for emerging pathogens, and including taxonomic division as a covariate.

3. Results

3.1. The validity of using automatically mined pathogen–location information in the EID2 database

3.1.1. Positive predictive value approach

If there was no direct evidence available from the NCBI Nucleotide database, a threshold (t) of number of papers from PubMed in which a pathogen name and MeSH term for a country co-occurred within each paper was used as evidence of pathogen presence. This was based upon a preliminary study in which papers had been stratified according to the pathogen and the continent to which they were linked via a MeSH term for a country. A putative association, identified in a publication by automated searches, was checked for accuracy to substantiate that the pathogen was found in hosts within a MeSH term country, after pathogen–MeSH term country combinations ($N = 21$) had been selected using random number generators. This allowed the calculation of the positive predictive value (PPV) of the test ($1 - ((1 - \mathrm{PPV})^t)$). On average, 20 out of 21 (95%) putative associations in single papers could be substantiated, giving a PPV of 0.95 (SE = 0.05) for evidence derived from a single publication. We therefore set a threshold for entry into EID2 of 5 papers providing evidence of the same association. As $1 - (0.05^5) > 0.999$, this threshold indicates that the PPV of entries into EID2, based on 5 or more publications, exceeds 99.9%. There was no evidence of pathogenic status ($N = 35$, $P = 0.393$) or taxonomic division ($N = 20$, $P = 1.00$) having a significant effect on the inferred relationship between a pathogen and a MeSH term country being true.

3.1.2. Binomial regression modelling approach

Having explored GAM results and a fifth order polynomial term within GLMs, a final binomial regression model included a $\log_{10}(n+1)$ transformation of the explanatory variable and a linear spline function with two breaks. The regression results suggest that the number of papers identified using PubMed is positively related to the odds of a pathogen–country interaction having been reported in the NCBI Nucleotide database, but the significance and characteristics of this relationship change dependent upon the number of PubMed papers (Fig. 1: overall results, df = 3, 3225, left-hand section of the line $P < 0.001$, middle section, $P < 0.001$, right-hand section, $P = 0.132$). The alternative model in which the number of papers from PubMed was recoded as a factor explained less of the residual deviance in the model (3989.1 as opposed to 3974.8, respectively), however, it too suggested two breaks points. Within this model, the odds of at least one nucleotide sequence describing a pathogen–country interaction increased significantly ($P < 0.001$) by 1.86 times (Confidence Interval (CI): 1.55–2.22) when there were between two and 12 PubMed papers in which a pathogen name and MeSH term for a country co-occurred, compared to when the interaction was described by only one PubMed paper. The odds of at least one nucleotide sequence increased significantly ($P < 0.001$) to 2.89 times (CI: 2.35–3.55) compared to one PubMed paper when the number of PubMed papers was more than 12.

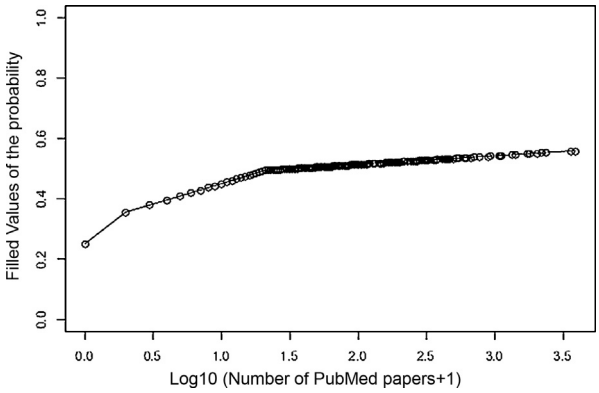

Fig. 1. The relationship between the number of papers identified using PubMed and the probability of a pathogen–country interaction having been reported in the NCBI Nucleotide database. The final generalised linear regression model with binomial errors and logit link function included a $\log_{10}(n+1)$ transformation of the explanatory variable (number of PubMed papers), and a linear spline function fitted with two breaks.

3.2. Summary of the EID2 database

3.2.1. Pathogen range within hosts

Of the final short-list of humans and domestic animal hosts (mammalian and avian livestock and pets) with which we have close contact in Europe, the EID2 described interactions connecting 47 hosts to 2597 pathogen species. Within the host species, 793 (30.5%) pathogens were bacteria species, 395 (15.2%) fungi, 705 (27.1%) helminths, 372 (14.3%) protozoa and 332 (12.8%) viruses. Pathogens generally affected very few different hosts, with the frequency of pathogen species to hosts characterised by a negative binomial distribution (median value = 1, mean value = 1.63, max value = 21; Fig. 2). Most (70.9%) pathogen species interacted with one host, 86.0% affected up to 2 hosts, and 98.6% affected six or fewer. The greatest proportion of pathogens (49.8%) affected humans only, 32.5% affected domestic animals and 17.8% affected both humans and animals. Of the human pathogens, 74.9% affected humans only and 25.1% affected humans and animals. Of the animal pathogens, 65.7% affected animals only and 34.3% affected humans and animals. The pathogens affecting the greatest number of hosts are presented in Table 2, including all those with at least seven hosts.

Table 2
The number of human or animal hosts affected by pathogen species, including taxonomic division. *E* or *NE* after pathogen names denote emerging or not emerging and *Z* or *NZ* denote zoonotic or non-zoonotic status according to Taylor et al. (2001) and Woolhouse and Gowtage-Sequeria (2005). NA denotes pathogens not included in these earlier studies.

Pathogen name	Pathogen type	Number of hosts	Pathogen name	Pathogen type	Number of hosts
Escherichia coli$_{E,Z}$	Bacteria	21	*Gongylonema pulchrum*$_{NE,Z}$	Helminth	9
Giardia intestinalis$_{E,Z}$	Protozoa	20	*Leptospira interrogans*$_{E,Z}$	Bacteria	9
Toxoplasma gondii$_{E,Z}$	Protozoa	18	Ovine Herpesvirus$_{NA}$	Virus	9
Anaplasma phagocytophilum$_{E,Z}$	Bacteria	15	Rotavirus A$_{E,Z}$	Virus	9
Cryptosporidium parvum$_{E,Z}$	Protozoa	14	*Clostridium perfringens*$_{NE,Z}$	Bacteria	8
Rabies virus$_{E,Z}$	Virus	13	Cowpox virus$_{NE,Z}$	Virus	8
Staphylococcus aureus$_{E,Z}$	Bacteria	13	*Enterococcus faecalis*$_{E,Z}$	Bacteria	8
Neospora caninum$_{NA}$	Protozoa	12	*Enterococcus faecium*$_{E,Z}$	Bacteria	8
Echinococcus granulosus$_{E,Z}$	Helminth	11	*Enterocytozoon bieneusi*$_{E,NZ}$	Fungi	8
Borna Disease virus$_{NE,Z}$	Virus	10	Hepatitis E virus$_{E,Z}$	Virus	8
Newcastle Disease virus$_{NE,Z}$	Virus	10	*Malassezia sympodialis*$_{NE,Z}$	Fungi	8
Pasteurella multocida$_{NE,Z}$	Bacteria	10	*Brachyspira pilosicoli*$_{NA}$	Bacteria	7
Trypanosoma cruzi$_{E,Z}$	Protozoa	10	Influenza A virus$_{E,Z}$	Virus	7
Babesia divergens$_{NE,Z}$	Protozoa	9	*Mecistocirrus digitatus*$_{NE,Z}$	Helminth	7
Chlamydophila psittaci$_{NE,Z}$	Bacteria	9	*Pneumocystis carinii*$_{E,Z}$	Fungi	7
Cryptosporidium muris$_{NA}$	Protozoa	9	*Saccharomyces cerevisiae*$_{NE,NZ}$	Fungi	7
Echinococcus canadensis$_{NA}$	Helminth	9	*Trichostrongylus colubriformis*$_{NE,Z}$	Helminth	7
Encephalitozoon cuniculi$_{E,Z}$	Fungi	9	West Nile virus$_{E,Z}$	Virus	7
Fasciola hepatica$_{NE,Z}$	Helminth	9			

Comparing pathogens which are emerging ($n = 169$) with those not emerging ($n = 1187$) (Table 3a) indicated that the odds of emergence differed by taxonomic division. The odds of emergence were 80% lower when the pathogens were helminths compared to bacteria, twice as high when they were protozoans and six times as high when they were viruses, both compared to bacteria. Further, the odds of pathogen emergence were 5 times higher if they had more than two compared to one host species associated with them, and more than one and a half times higher if they were zoonotic compared to non-zoonotic. Pathogenicity did not significantly affect the odds of a pathogen being emerging.

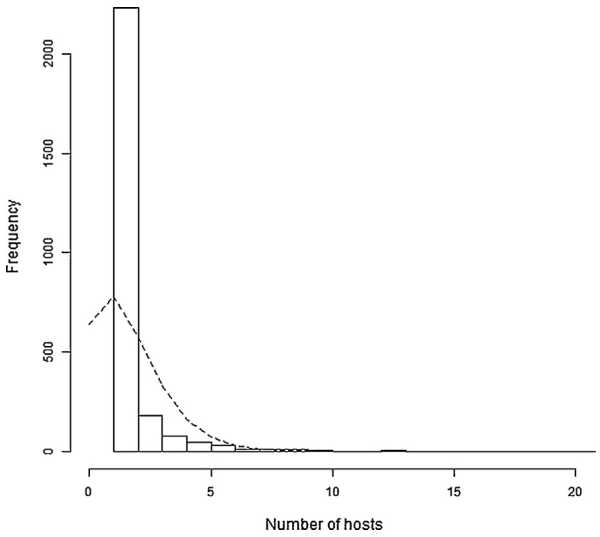

Fig. 2. Histogram of the frequency of pathogen species affecting humans and domestic animal hosts, including a theoretical density function based on empirical estimates of the parameters of the negative binomial distribution (size = 4.805, mean(mu) = 1.626).

Comparing pathogens which are zoonotic ($n = 829$) with those non-zoonotic ($n = 527$) (Table 3b) indicated that the odds of a pathogen being zoonotic differed by taxonomic division. The odds of a pathogen being zoonotic were 16 times higher if they were helminth pathogens, more than twice as high if they were protozoans and nearly four

Table 3
Logistic regression models for the odds of pathogens being emerging (a) or zoonotic (b), including Adjusted Odds Ratios (AOR), 95% Confidence Limits (95% CL) and *P* values for the statistical significance of a difference in each factor-level relative to the baseline.

Attribute	AOR	95% CL		*P* value
		Lower	Upper	
(a)				
Number of hosts species				
1	Baseline	–	–	–
2	1.41	0.75	2.63	0.285
>2	4.82	3.00	7.74	<0.001
Taxonomic division				
Bacteria	Baseline	–	–	
Fungi	1.02	0.59	1.75	0.954
Helminths	0.18	0.09	0.37	<0.001
Protozoa	2.06	1.00	4.21	0.049
Viruses	6.16	3.94	9.63	<0.001
Zoonotic status				
Non-zoonotic	Baseline	–	–	–
Zoonotic	1.64	1.07	2.53	0.023
Hosmer–Lemeshow goodness-of-fit test		$P = 0.47$		
(b)				
Number of hosts species				
2	Baseline	–	–	–
1	0.28	0.17	0.46	<0.001
>2	3.09	1.45	6.58	0.003
Taxonomic division				
Bacteria	Baseline	–	–	–
Fungi	0.75	0.55	1.02	0.063
Helminths	16.35	9.04	29.59	<0.001
Protozoa	2.29	1.17	4.50	0.016
Viruses	3.77	2.56	5.56	<0.001
Hosmer–Lemeshow goodness-of-fit test		$P = 0.54$		

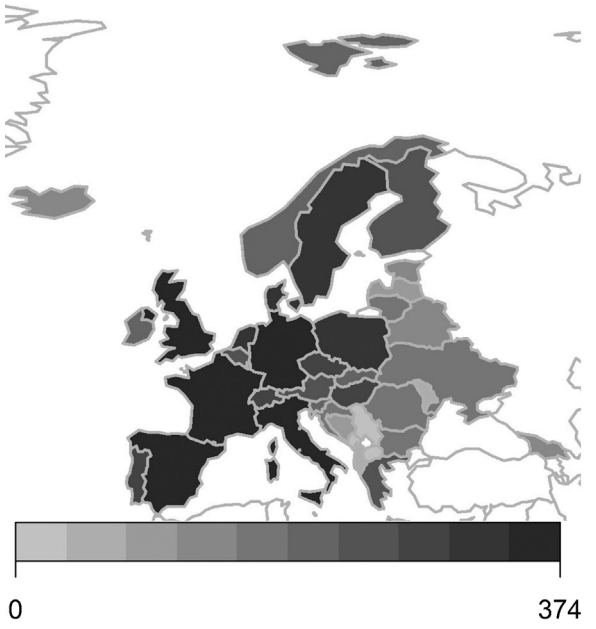

Fig. 3. Map of Europe showing number of pathogen species in each country. Scale depicts the number of species.

times higher if they were viruses, all compared to bacteria. Further, the odds of pathogens being zoonotic were more than three times higher if they had more than two compared to two host species associated with them, and 70% lower if they had only one host compared to two host species associated with them. Pathogenicity did not significantly affect the odds of a pathogen being zoonotic. Emerging status caused the model to become unstable and it was therefore not included within the final results; when forced in, it increased the odds of pathogens being zoonotic.

3.2.2. Host range

The pathogens of humans and domestic animal hosts are characterised by 4223 host–pathogen interactions, with the greatest number found in (decreasing order): humans, sheep/goats, cattle, small mammals, pigs, dogs, equids, cats, chickens, other edible birds, deer, exotic mammals, ducks and caged birds (Table 4). The breakdown of pathogen species by taxonomic division is given in Table 4. Pathogens of deer were most likely to cause significant clinical effects in one of their hosts as opposed to being non-pathogenic in any host, followed by those of (descending): cattle, pigs, dogs, humans, exotic mammals, cats, ducks, caged birds, equids, sheep/goats, other edible birds, small mammals and chickens.

3.2.3. Differences in the spatial representation of pathogen species

The number of pathogen species varied by European country (Fig. 3). In all host groups except cattle, small mammals, pigs, deer and ducks, the number of pathogens reported was greater in the rest of the world than in Europe (Table 4).

Comparing pathogens in Europe ($n = 601$) as opposed to in the rest of the world ($n = 755$) (Table 5) indicated that the odds of a pathogen being found in Europe differed by taxonomic division. The odds of a pathogen being found in Europe were 80% lower when the pathogen was a helminth, 50% lower when it was a protozoa, and 60% lower when it was a virus, all compared to bacteria. Further, the odds of pathogens being found in Europe were 3 times or nearly 7 times higher if they had 2 or more than 2 hosts compared to one host species associated with them, respectively. The odds of a pathogen being found in Europe compared to the rest of the world were more than 3 and a half times higher if they were emerging compared to not emerging. Zoonotic status did not significantly affect the odds of a pathogen being found in Europe.

4. Discussion

The EID2 database described here utilises the individual reports of host–pathogen interactions provided by genome, gene and transcript sequence recording resources. This study demonstrates the vast amount of hitherto untapped information which can be used for approaches to examine the characteristics and drivers of pathogens, and particularly those affecting multiple hosts. It illustrates that by using such methods we can examine differences in surveillance of pathogens, but also begin to untangle the vast networks of hosts in which they occur. An approach such as this, which embraces multiple and unconventional data sources, could lead to significant advances in the quantification of the global burden of disease, in international biosurveillance and in exploring disease outbreaks, emergence and evolution, as suggested recently by Hay et al. (2013). There are biases inherent in the use of sequence databank resources, however these are likely to become less pronounced over time, with probable improvements in both the coverage and accuracy of information; this contrasts with the use of information from text books and other printed media. The biases in sequence information include likely under-reporting of pathogens for various reasons: they might not cause clinical infection and thus reporting would not have tangible benefits, they might not cause a notifiable disease or their geographical distribution or host range might have changed so that they would not be sought in host species; they might be prevalent in countries in which surveillance resources are restricted meaning a lack of identification but also submissions to sequence databases would not occur; and for some pathogen groups such as fungi, it might be difficult or unnecessary in clinical practice to identify species, when a cure can be secured without extra effort. Another bias might be in the hosts in which pathogens are reported: it is likely that some species such as humans will be relatively over-represented whereas little may be known or reported on the pathogens of wildlife; and pathogens may be reported from research on laboratory animals rather than after infection of natural hosts. Further, biological material for certain pathogens may be easier to obtain than for others, either because they are physically difficult or costly to isolate within a host.

Table 4
Number of human or animal host and pathogen species interactions (*N*) (summarised by host groups) in the EID2, including the percentage in each taxonomic division and the percentage in Europe compared to the rest of the world.

Host group	Species included in host group	*N*	% of pathogens in taxonomic division					% of pathogens	
			Bacteria	Fungi	Helminths	Protozoa	Viruses	European	Non-European
Humans	*Homo sapiens*	1752	39.4	22.1	20.1	4.7	13.6	45.1	54.9
Sheep/goats	*Capra hircus, Ovis aries, Ovis aries musimon*	371	18.9	1.3	42.3	25.9	11.6	48.2	51.8
Cattle	*Bison bison, Bison bonasus, Bos indicus, Bos taurus*	353	26.9	2.3	34.6	25.2	11.0	54.7	45.3
Small mammals	*Cavia porcellus, Chinchilla lanigera, Cricetus cricetus, Meriones unguiculatus, Mesocricetus auratus, Mus musculus, Mustela putorius furo, Oryctolagus cuniculus, Rattus norvegicus, Rattus rattus, Rhombomys opimus*	346	25.4	6.1	19.1	36.4	13.0	58.1	41.9
Pigs	*Sus scrofa, Sus scrofa domesticus*	330	28.5	2.1	41.5	16.7	11.2	52.7	47.3
Dogs	*Canis lupus familiaris*	228	20.6	3.9	54.8	16.2	4.4	43.4	56.6
Equids	*Equus asinus, Equus caballus*	164	20.7	6.7	37.2	20.1	15.2	46.3	53.7
Cats	*Felis catus*	161	12.4	3.7	60.2	15.5	8.1	34.2	65.8
Chickens	*Gallus gallus*	136	31.6	2.9	36.0	16.9	12.5	47.1	52.9
Other edible birds	*Anser anser, Columba livia, Lagopus lagopus scotica, Meleagris gallopavo, Numida meleagris, Phasianus colchicus, Struthio camelus*	123	17.1	2.4	30.9	32.5	17.1	43.1	56.9
Deer	*Capreolus capreolus, Cervus elaphus, Dama dama, Rangifer tarandus*	111	17.1	0.9	44.1	29.7	8.1	74.8	25.2
Exotic mammals	*Camelus dromedarius, Lama glama, Lama pacos*	78	11.5	2.6	41.0	30.8	14.1	42.3	57.7
Ducks	*Anas platyrhynchos*	40	45.0	0.0	10.0	25.0	20.0	65.0	35.0
Caged birds	*Agapornis personata, Agapornis roseicollis, Melopsittacus undulatus, Nymphicus hollandicus, Pavo cristatus, Serinus canaria*	30	13.3	3.3	16.7	30.0	36.7	50.0	50.0

Table 5
Logistic regression model for the odds of pathogens being found in Europe as opposed to the rest of the world, including Adjusted Odds Ratios (AOR), 95% Confidence Limits (95% CL) and *P* values for the statistical significance of a difference in each factor-level relative to the baseline.

Attribute	AOR	95% CL		*P* value
		Lower	Upper	
Number of hosts species				
1	Baseline	–	–	–
2	2.97	1.97	4.46	<0.001
>2	6.66	4.44	10.01	<0.001
Taxonomic division				
Bacteria	Baseline	–	–	–
Fungi	0.94	0.70	1.26	0.659
Helminths	0.17	0.11	0.25	<0.001
Protozoa	0.48	0.26	0.91	0.025
Viruses	0.39	0.27	0.59	<0.001
Emerging status				
Not emerging	Baseline	–	–	–
Emerging	3.68	2.43	5.58	<0.001
Hosmer–Lemeshow goodness-of-fit test		*P* = 0.22		

4.1. The validity of using automatically mined pathogen–location information in the EID2 database

The initial analyses within this study validate the use of semi-automatic data-mining processes within the EID2 to ascertain where pathogens occur at the country level, given consideration of sample sizes. Future developments of the database will include undertaking similar tests for automatic assignation of host–pathogen interactions, and increasing the spatial resolution of pathogen–location information.

4.2. Summary of the EID2 database

4.2.1. Pathogen range within hosts

Considering the proportion of pathogens found within certain taxonomic divisions, the results for bacteria, fungi and helminths were largely similar for this study in comparison to previous work examining human, livestock and domestic carnivore populations (Cleaveland et al., 2001). In this study compared to Cleaveland et al. (2001), the proportions of protozoan species were higher, however, and viruses were slightly lower, possibly due to the additional inclusion of small mammal, domestic bird and exotic (domestic animal) host species. This study also examined pathogens not causing clinical infection or with unknown aetiologies, where the other study used only clinical disease.

Within this study, the proportion of pathogens affecting more than one host was much lower than for Cleaveland et al. (2001) (29.1% compared to 62.7%), perhaps due to the inclusion of non-clinical infection but also because nucleotide sequences are more likely to be sent into a sequence databank if they have been found for the first time in a novel host or unusual environment, potentially skewing the distribution of data on pathogens affecting host species. Cleaveland et al. (2001) reported 39.1% of human pathogens infecting domestic animal hosts; a higher proportion than found within this study (25.1%) because given the relative investment in healthcare systems, information on non-clinical disease is more likely to be ascertained for human than animal hosts. This would further explain the slightly lower proportion of livestock pathogens affecting humans in this than the Cleaveland study (34.3% and 39.4%, respectively). That said, a conclusion that fewer multi-host pathogens cause disease in humans than domestic animals was found in both studies, and postulated upon by Cleaveland et al. (2001). We acknowledge that these comparisons suggest that domestic animal infections may be currently under-represented in the EID2, but emphasise that information should become more complete over time, as more sequences are uploaded and papers published.

The results of comparison of emerging and not emerging pathogens have corroborated previous work: pathogen emergence is associated with the number of host species, with the odds of emergence increasing with greater numbers (Cleaveland et al., 2001; Woolhouse and Gowtage-Sequeria, 2005); the odds of helminths being emerging were lower and the odds of viruses and protozoa were higher (Cleaveland et al., 2001; Dobson and Foufopoulos, 2001; Taylor et al., 2001), though a reduced odds of emergence for fungi (Cleaveland et al., 2001; Taylor et al., 2001) was not observed in this study, perhaps as the comparison was relative to bacteria; and the odds of pathogens being emerging compared to not emerging were higher for zoonotic as opposed to non-zoonotic pathogen species (Cleaveland et al., 2001; Taylor et al., 2001; Woolhouse and Gowtage-Sequeria, 2005; Jones et al., 2008).

That zoonotic compared to non-zoonotic pathogens were influenced by number of host species reflects the capacity of a pathogen to spread or transmit within host populations; disease transmission is more likely given a network of hosts which may be closely phylogenically related or which interact closely as with domesticated animal populations (Janes et al., 2012). That a statistically significant difference was found in the odds of pathogens being zoonotic compared to non-zoonotic when comparing pathogens associated with two as opposed to one host species, reflects likely gaps in the host data within the EID2 database.

4.2.2. Host range and the spatial representation of pathogen species

Many of the results in this study for host range of pathogens and the countries in which each pathogen species occurred may be influenced by sampling biases in surveillance. For instance, the number of host–pathogen interactions identified within host groups reflects differences in healthcare, politics and the funding of disease surveillance (Daszak et al., 2000). Patterns in the proportion of clinical effects within host groups likewise could be due to biases in non-clinical pathogens not being isolated within certain host species, for instance deer, or being less likely to be isolated in, for example, developing world countries compared to Europe. Spatial sampling biases are reflected in the map illustration in which eastern European countries had much fewer numbers of pathogen species compared to western European; pathogens isolations from eastern Europe before relatively recent country boundary

changes may not have been included in EID2 data. Further, comparison of pathogens in Europe and the rest of the world suggested differences in sampling effort between taxonomic divisions, when fewer host species were associated with pathogens, and when pathogen species had been awarded a politically important status: they were emerging. This final result might be surprising given the relatively recent plethora of publications and funding on emerging pathogens (Taylor et al., 2001; Woolhouse and Gowtage-Sequeria, 2005; Murphy, 2008; Woolhouse, 2008), much of which have come from North America (Daszak et al., 2000, 2001; Dobson and Foufopoulos, 2001; Jones et al., 2008). It might also, however, reflect the sheer quantity of pathogen isolations undertaken in areas outside Europe.

Future work aims to expand the information held within the EID2, to build a more comprehensive list of livestock and other pathogens. Specific improvements will include: greater spatially detailed information for pathogens; improvements to the database's ability to handle records with badly defined host species; the addition of further environmental data to produce better models to explain pathogen distributions and predict them in the future, given climate change; and information to allow users to work at the level of diseases, rather than individual pathogens or groups of pathogens.

5. Conclusion

Previous studies and reviews suggest that in order to provide early warning systems for disease emergence and to study zoonotic pathogens, a united effort is needed to collate information on their complex epidemiologies (Daszak et al., 2000; Cleaveland et al., 2001; Taylor et al., 2001; Woolhouse and Gowtage-Sequeria, 2005; Wolfe et al., 2007; Murphy, 2008; Janes et al., 2012). It is hoped that the work described within this study goes some way to demonstrating that much of the necessary information may potentially already be available and can be harnessed using semi-automated methodologies such as that provided by the EID2; by being more methodical within our unified efforts, we can build 'one health' disease surveillance systems, as suggested by Hay et al. (2013). Potentially, the EID2 will benefit the veterinary and human health communities by providing greater lead-time for pathogen surveillance or for control measure design, and it will help inform clinicians about pathogens driving clinical disease; their origins, a temporal and spatial indication of recent disease outbreaks and links to publications which describe them.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgements

The authors would like to thank Helen Roberts (UK Department for Environment, Food and Rural Affairs) and John Stephenson (formerly from the Health Protection Agency) for their input during the development of the EID2 database, and Sally Eagle for help with the piecewise regression modelling. This work was funded by NERC grant [NE/G002827/1] through the ERA-ENVHEALTH network, awarded to MB, and by a BBSRC Strategic Tools and Resources Development Fund grant [BB/K003798/1], awarded to MB and ADR.

References

Cleaveland, S., Laurenson, M.K., Taylor, L.H., 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos. Trans. R. Soc. B 356, 991–999.

Daszak, P., Cunningham, A.A., Hyatt, A.D., 2000. Wildlife ecology – emerging infectious diseases of wildlife – threats to biodiversity and human health. Science 287, 443–449.

Daszak, P., Cunningham, A.A., Hyatt, A.D., 2001. Anthropogenic environmental change and the emergence of infectious diseases in wildlife. Acta Trop. 78, 103–116.

Dobson, A., Foufopoulos, J., 2001. Emerging infectious pathogens of wildlife. Philos. Trans. R. Soc. B 356, 1001–1012.

Hay, S.I., Battle, K.E., Pigott, D.M., Smith, D.L., Moyes, C.L., Bhatt, S., Brownstein, J.S., Collier, N., Myers, M.F., George, D.B., Gething, P.W., 2013. Global mapping of infectious disease. Philos. Trans. R. Soc. B 368 (1614), 20120250.

Janes, C.R., Corbett, K.K., Jones, J.H., Trostle, J., 2012. Emerging infectious diseases: the role of social sciences. Lancet 380, 1884–1886.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990–994.

McIntyre, K.M., Setzkorn, C., Baylis, M., Waret-Szkuta, A., Caminade, C., Morse, A.P., Akin, S.M., Huynen, M., Martens, P., Morand, S., 2010. Impact of climate change on human and animal health. Vet. Rec. 167, 586–658.

Moore, G.E., Ward, M.P., Dhariwal, J., Wu, C.C., Glickman, N.W., Lewis, H.B., Glickman, L.T., 2004a. Development of a national companion animal syndromic surveillance system for bioterrorism. In: 2nd International Conference on the Applications of GIS and Spatial Analysis to Veterinary Science (GISVET 04), Univ. Guelph, Ontario, Canada.

Moore, G.E., Ward, M.P., Dhariwal, J., Wu, C.C., Glickman, N.W., Lewis, H.B., Glickman, L.T., 2004b. Use of a primary care veterinary medical database for surveillance of syndromes and diseases in dogs and cats. J. Vet. Intern. Med. 18, 386–386.

Murphy, F.A., 2008. Emerging zoonoses: the challenge for public health and biodefense. Prev. Vet. Med. 86, 216–223.

National Center for Biotechnology Information, 2012a. US National Library of Medicine, Bethesda, Maryland, US. The NCBI Medical Subject Headings (MeSH) Database Homepage, http://www.ncbi.nlm.nih.gov/mesh

National Center for Biotechnology Information, 2012b. US National Library of Medicine, Bethesda, Maryland, US. The NCBI Nucleotide Database Homepage, http://www.ncbi.nlm.nih.gov/nuccore

National Center for Biotechnology Information, 2012c. US National Library of Medicine, Bethesda, Maryland, US. The NCBI Taxonomy Database Homepage, http://www.ncbi.nlm.nih.gov/Taxonomy/

National Center for Biotechnology Information, 2012d. US National Library of Medicine, Bethesda, Maryland, US. PubMed, http://www.ncbi.nlm.nih.gov/pubmed/

Paiba, G.A., Roberts, S.R., Houston, C.W., Williams, E.C., Smith, L.H., Gibbens, J.C., Holdship, S., Lysons, R., 2007. UK surveillance: provision of quality assured information from combined datasets. Prev. Vet. Med. 81, 117–134.

Radford, A., Tierney, A., Coyne, K.P., Gaskell, R.M., Noble, P.J., Dawson, S., Setzkorn, C., Jones, P.H., Buchan, I.E., Newton, J.R., Bryan, J.G.E., 2010. Developing a network for small animal disease surveillance. Vet. Rec. 167, 472–474.

Tam, C.C., Rodrigues, L.C., Viviani, L., Dodds, J.P., Evans, M.R., Hunter, P.R., Gray, J.J., Letley, L.H., Rait, G., Tompkins, D.S., O'Brien, S.J., Comm, I.I.D.S.E., 2012. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. Gut 61, 69–77.

Taylor, L.H., Latham, S.M., Woolhouse, M.E.J., 2001. Risk factors for human disease emergence. Philos. Trans. R. Soc. B 356, 983–989.

Thrusfield, M., 2007. Veterinary Epidemiology. Blackwell Science Ltd., pp. 316–318.

University of Liverpool (2011), 2011. The ENHanCEd Infectious Diseases database (EID2). National Centre for Zoonosis Research, www.zoonosis.ac.uk/eid2

Wolfe, N.D., Dunavan, C.P., Diamond, J., 2007. Origins of major human infectious diseases. Nature 447, 279–283.

Woolhouse, M.E.J., 2008. Epidemiology – emerging diseases go global. Nature 451, 898–899.

Woolhouse, M.E.J., Gowtage-Sequeria, S., 2005. Host range and emerging and reemerging pathogens. Emerg. Infect. Dis. 11, 1842–1847.